International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

Text-Based Domain and Image Categorization of Google Search Engine using Conceptual Clustering

Divya Soni[1], Preetesh Purohit[2]

[1]Student, Swami Vivekanand College of Engineering, Indore, India
[2]Associate Professor, Swami Vivekanand College of Engineering, Indore, India

---

Abstract - The search engine like google provides a record of results that show a list of ranked output. The ranked output is not considered user-relevant. Query suggestion system is an option for this problem. But still, users are always forced to shift through the long ordered list of documents retrieved by the search engines. However, categorization is one option to solve this problem. The proposed system performs categorization based on several relevant conceptual information and predictive modelling. Proposed work acquaints an impressive approach that takes the conceptual preferences of users to get topics of the snippet with the help of analyzed data. To achieve this goal, an online technique is introduced that analyzes web-content of the generated results to find appropriate concepts and uses it to recognize related results. A personalized conceptual clustering algorithm is also used to generate a decision tree of clusters of the query data. With the help of this tree, the proposed system can identify classes for the web pages easily which provide relevant topical results to the user.

Key Words: Web Mining, Text Mining, Clustering, Web Page Recommendation, Search Engine, Query Logs, Domain Categorization.

1. INTRODUCTION

Search engines are web portals for finding knowledge on world wide web. Search engines index a large portion of the web and store the information in databases. Unlike a web directory, the search engine carries operations automatically. The underlying technique of search engines is information retrieval. The obvious question in information retrieval is how to find the relevant documents for each user query. This is the main issue that tries to address [1].

Knowledge mining is a procedure of mining usable information from the crude information. The information is a term that is principal to the information that is required by way of a data miner or utility. Extraction of the similar text from a raw set of text is the generation of text data mining. Clearly, textual information does not follow any similar pattern in text data, an unstructured technique and labelling of text data is tricky initiation. Therefore, several applications make use of the classification and clustering techniques for categorizing knowledge.

Finding high quality labelled document is crucial for any search engine since it can be used to estimate the search engine quality and efficiency. The basic idea of proposed work is established on concepts and its similarities extracted from the search queries submitted by the user, the web content, and the click data. To identify user preference, the personalized clustering technique is applied which exploit these click-through data: A user clicks a link in search results if the result has relevant web content in which the user interested. Moreover, click data can be gathered easily without the effective excess load on users, thus providing a low-cost means to capture user's interest [2][3][4].

2. BACKGROUND

World Wide Web contains a huge amount of knowledge and data, some of the knowledge is distributed using the contents of web pages and some of the information is not directly gained from the direct web. Therefore, in order to recover the meaningful and essential knowledge from the web, a domain known as web mining is introduced. Web mining utilizes the web data and the mining techniques. The web mining is divided in three main key domains according to the kind of data processing. The content of web page is analyzed using content mining, web access log is analyzed under web usages mining and the connectivity of the web pages are analyzed under the structure mining. Recommendation systems are the most essential application of web mining and most of the recommender systems typically generate a list of recommendations in two ways - content-based filtering or through collaborative. Recommendation system algorithm based on content filtering [5] is the algorithm that works with content based profiles that are created at the beginning from the users' preferences. This profile contains information about a user and his interesting topics. Topics are based on how many times a user retrieve the same results for the related query. In this process, the search engine analyzes the results looks for similarities that were previously patently ranked by the user with the results he did not ranked. Those results that are chiefly similar to the patently ranked ones, will be recommended to the user. Collaborative filtering Algorithm [6][7] recommendation system was one of the most popular researched technique of recommendation systems since that technique was described and stated by H. Varian and P. Resnick in 1997. [8][9] collaborative filtering is based on community data analysis which find the users from their shared appreciations in the community [10]. If multiple users have similar or nearly same ranked results in common, then those users have similar interest in that result [11].

System uses such users to build a set or a neighborhood. System uses such users to build a set or a neighborhood. A user gets only those results which are not ranked before in his recommendations list. But gets those results that were previously already ranked by any user in his/her set or neighborhood. The locations of the users can be used by the search engines to construct a set or neighborhood that can be used for the categorization of the search results [12].

3. RELATED WORK

Based on implicit observation, there are two fundamental types of categorization techniques: document-based and concept-based [13] [14]. Techniques of document-based approach uses overall click through data to identify user document preferences. These preferences are used to analyze functions of ranking that optimize clicking and browsing preferences of a user on the retrieved data. And concept-based techniques focus on determining user thematic choice from the content of the users' browsed data

Joachim's method [15] proposed optimization of search engine using click through data that works on ranking systems. A user would perceive the search results from top to bottom because of click through data are used to extract user clicking preferences. A ranking based SVM algorithm uses the click preference [16] to teach a ranker which best fits preferences of the users. Q. Tan et al [17] proposed RSVM algorithm which works on the ranking SVM algorithm with a co-training framework for the available ranking system. In 2012 Samuel Ieong et al proposed an undeniable proof of the existence of domain bias that uses Shannon entropy to measure the view and visit distribution of domains. When entropy increases, it means user visits becoming more miscellaneous, while decreasing entropy is a sign of user visits becoming foreseeable and propounds the formation of domain preferences [3]. In 2012 Bo Geng, Linjun Yang et al proposed an algorithm based on a regularization called ranking adaptation SVM (RASVM). RASVM uses an existing ranking model to a new domain, which reduces training cost and labelled data while the performance still guaranteed [4]. In 2004 F. Liu et al proposed "Personalized Web Search for Improving Retrieval Effectiveness" that shows the preferences for independent model for long-term and short-term. The Google Directory used to determine long-term preferences, while the short-term preferences are determined from the user's data preferences that are most repeatedly browsed data [18]. In 2006 L. Deng et al proposed a spying technique that merged with a novel voting technique to find preferences of user data from the clicking information, this technique spying on search data. [19]. In 2014 R.A. B-Yates et al proposed an advanced clustering technique which works on search queries. This technique groups relevant search queries according to their semantics. This technique builds a vector representation "Q" for a search query "q" and the created vector "Q" contains terms from the clicked documents of "q". Cosine symmetry method is also applied to the query vectors to find similar queries [20]. In 2013 S. Chuang et al proposed a method that creates user profiles based on search results of user history data and a predefined taxonomy. The user profiles that contain user-preferred categories from ODP are then give categorized search results [21]. In 2007 Z. Dou et al proposed an approach that developing users' profiles based on the ODP taxonomy to produce labeled search results. A dispersive algorithm proposed to maintain the scores of the user interest on the ODP categories based on ongoing behavior of users [22]. A similar adaptive strategy based on behavior of users also proposed. Instead of using ODP as the taxonomy, this uses Google Directory-3 as the predefined taxonomy to build a user profile. In 2016 B. Koester proposed a framework to improve search engine results with contexts and concept hierarchies, which supports in mining the conceptual preferences of a user from the click through data of the user, resulting from Web search. The framework build an expanded set of concept based preferences of users from extracted concepts from the search results and the preferences click through data. Using these preferences, it creates a user profile which is represented as a conceptual ontology tree by the concept based user profile (CUP) [23]. Finally, the CUP is given as an input to a support vector machine (SVM) to learn the concept preferences of the user to re-rank the search results. In 2017 Jiawei Liu et al proposed a query suggestion method based on random walk and topic concepts (QuS-RWTC). The method is based on the query log data and suggestions from other mature search engines, which could make the suggestions more comprehensive and obtain a higher coverage. In addition, the paper further executes procedures of topic concepts to reorder the candidate queries, which make the suggestions more accurate, since they are more satisfied to the user's initial intention [24]. In 2016 Asri Maspupah and Saiful Akbar proposed search application framework in which developer do not have to start development from the scratch. Instead, he/she can develop the application by customizing the framework in accordance with their needs. The research focuses on how to group similar aspects that are available in various search engines. The development process of search application framework is done by generalizing process and identifying the variability of the similar processes; and designing the scheme of the framework using design pattern. The proposed framework has been tested by implementing several search engine applications with different search methods using the framework. The implementation has been applied as a mean to evaluate the application of the framework. The contribution of this research is a framework that can be utilized as a tool to support the development of search engine application [25].

In 2017 Hoang Giang et al proposed an efficient document modelling framework that can be applied in real-world applications, such as search engines and web recommendation systems. The document modelling framework works on topic models and vector space models with some improved factors to process documents prior to

modelling and improve the efficiency [26]. Finally, a similitude is drawn between some previous search engine categorization research.

Table-1: Previous research methods and limitations

Research	Method Used	Advantage	Disadvantage
Domain Bias in Web Search	Shannon entropy	Provide indisputable proof of the existence of domain bias	Approach is limited in scale due to the reliance of human labels
Ranking Model Adaptation for Domain-Specific Search	Ranking adaption with domain specific feature	Small number of samples need to be labeled, Less computation cost	Limits their query answering spectrum to narrow queries over specific domain knowledge
A Vertical Search Engine – Based On Domain Classifier	Content analysis technique	Provide relevant results to the user	Limited to only two domains Medical and finance

4. PROPOSED METHODOLOGY

The information available on web can be available in different formats and can be utilized with different kinds of applications. Among them the recommendation system is an essential area of web usage mining and recommendation system design this section provides methodology and methods that can help to archive our goal

4.1 Web Access Log

A server log is also known as the web access log file created and maintained by web server to manage the activities. A general web access log maintains a record of clients' or users' generated page requests. The W3C is a norm and default web log format for server log. The entries which are not present already are merged at the end of the file. This contains information about, unique address of a user, request date - time stamp, request functions, request protocol, response code, HTTP code, operating system(OS), OS version, browser and browser information sent, user agent, information received and referrer etc. Such files are not accessible to Internet users directly, only webmasters or administrator can get access the web access data due to privacy and security measures. Analysis of these files may use to find traffic patterns and other applications.

4.2 Concept Extraction

The concept extraction method composed of two basic steps.

Step 1: To extract the concepts making use of the web-snippets returned back from the search engine for a user query.

Step 2: To obtain the relation between the extracted concepts.

When a user submits a query to the search engine, it returns a set of web-snippets back to the user to identify the items that are relevant to the user. If the returned web-snippets, for a particular query contain keywords or phrases that appear frequently, then those keywords or phrases will be considered as an important concept. The extracted concepts will relate to the user query, as it lies in close proximity with the user query among the top documents returned.

5. PROPOSED SYSTEM

Personalized conceptual clustering algorithm is an efficient algorithm to generate new different query clusters from the ambiguous queries. In the clustering process, the user profile of concept preference is used to achieve personalized effect. In contrast to agglomerative clustering algorithm, which represents the same queries submitted from different users by one query node, we need to consider the same queries submitted by multiple users individually to obtain personalization effect. In other words, if two given queries, whether they are identical or not, mean different things to two different users, they should not be merged together because they refer to two different concept sets of the two different users. Therefore, each individual query of each user treated as an individual vertex in the graph by assigning labels to each query with an identification token of each user. For example, we can see that a query "Apple" submitted by users "UserA" and "UserB" become two vertices "Apple UserA" and "Apple UserB". For both users there may be different results from their interest. If UserA is interested in the "Apple iPhone" which was recorded in the preference profile of user concepts, a link between the concept "Apple iPhone" and the query "Apple UserA" would be created. On the other hand, if UserB wants search results from apple fruit a link between the concept "fruit" and "Apple UserB" would be created. This process generates a personalized graph and our preparatory experiments manifested that if we apply algorithm directly on generated graph, the generated query clusters will merge queries instantly from different users together, thus the system loses the personalization effect.

It is found that similar queries submitted by multiple users and which also have multiple meanings, tend to contain some general concept nodes such as "details" in common. e.g. "Apple UserA" and "Apple UserB" both have connection to the "details" concept node. Thus, these nodes of queries in the first few steps will eventually be merged and because

number of multiple queries from multiple users to be merged together in subsequent iterations. "Apple iPhone", "fruit" and "details" concept nodes will be merged in the next iteration. When the clustering algorithm goes further, queries across users will be further clustered together. Therefore, there is no personalization effect in the resulting query clusters at the end.

To overcome this problem, we are dividing clustering process into two processes. In the preparative clustering process, we apply a clustering algorithm to cluster user queries which is similar to the COBWEB clustering algorithm. The algorithm would not merge similar queries of multiple users initially. After initial clustering process we obtain all the clusters and then community merging process is applied to merge the clusters of the similar queries from the multiple users.

Proposed model is an effective solution to generate text based categories for domain and images on known and unknown URLs that may be frequently visited or not by the user. Figure 1 shows the architecture of proposed system that input user query and search results and returns user relevant results. Results retrieved from search engine are used to extract concepts and click though data. Query logs are also used to create user profile. User profile contains analyzed data of query logs, concept tree and click though data. Conceptual clustering algorithm processes on it and returns relevant data to the user.

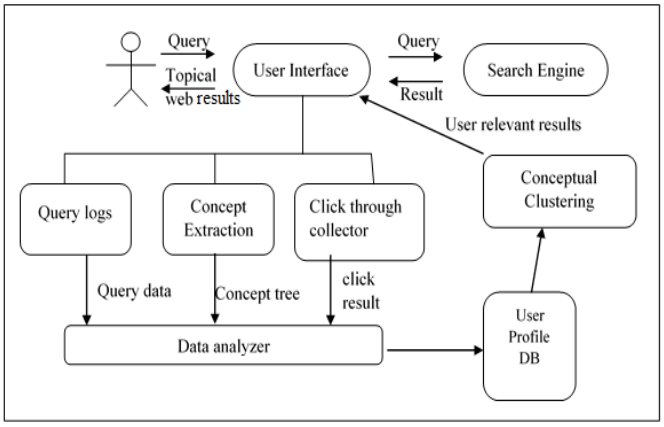

Fig -1: Proposed Model

Algorithm

Step 1: input unknown web text.

Step 2: use terms and phrase to generate user concepts.

Step 3: build concept tree based on detail node of each user.

Step 4: merge common concept node with corresponding user node.

Step 5: analyze the concept tree with query data and click through results.

Step 6: create concept profile for each user.

Step 7: apply clustering algorithm.

Step 8: merge the clusters of the similar queries of the multiple users.

Step 9: extract the relevant topics from the clusters.

Step 10: output categorized results.

6. CONCLUSION

Thousands of domains are there, finding the right one from them is quite difficult process. Searching process takes too much time and many filtrations to search out the right one from them. Search queries may be ambiguous; we have studied dominant techniques for search engines to provide relevant search results on semantically respective queries in order to help users formulate more effective results to meet their diversified needs. We also studied concept extraction from the user experiences that helps to find relevant data. Clustering is the one solution to find relevant data from the web and we will use conceptual clustering technique that is able to obtain topic vice web pages and images for individual users based on their conceptual profiles. Proposed work uses query suggestion logs, click through data and the concept relationship graph mined from web-snippets that can be captured at the back end and as such do not add extra burden to users. By knowing the intended information requirement of the user implicitly, the performance of the search engine can be improved. A user behavior is needed to recognize the information need of the users that is very useful to give relevant results.

REFERENCES

[1] S.M. Beitzel, E.C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. "Hourly Analysis of a Very Large Topically Categorized Web Query Log," Proc. 27th Ann. Int'l ACM SIGIR Conf. (SIGIR), 2004.

[2] E. Agichtein, E. Brill, and S. Dumais, "Learning User Interaction Models for Predicting Web Search Result Preferences," Proc. 29th Ann. Int'l ACM SIGIR Conf. (SIGIR), 2006.

[3] Samuel Ieong, Nina Mishra, Eldar Sadikov, Li Zhang, "Domain Bias in Web Search", ACM New York, NY, USA ©2012.

[4] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, "Ranking Model Adaptation for Domain-Specific Search", IEEE Transactions on Systems knowledge and data engineering vol. 24, no. 4, April 2012.

[5] Z. Zhang and O. Nasraoui, "Mining Search Engine Query Logs for Query Recommendation," Proc. 15th Int'l World Wide Web Conf. (WWW), 2006.

[6] Umajancy. S, Dr. Antony Selvadoss Thanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[7] Hien Nguyen, Eugene Santos, and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6, November 2011.

[8] Miloš Radovanović, Mirjana Ivanović, "Text Mining: Approaches and Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008.

[9] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[10] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime G. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", ACM SIGIR Conference on Research and Development in information Retrieval, July 1999.

[11] Andreas Hotho, Andreas Nurnberger, Gerhard Paab, Fraunhofer AiS, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005

[12] Weiguo Fan, Linda Wallace, Stephanie RichZhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005

[13] Daniel A. Keim, " Information Visualization and Visual Data Mining", IEEE transactions on visualization and computer graphics, vol. 7, no. 1, January-march 2002.

[14] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[15] T. Joachims, "Optimizing Search Engines Using Click through Data," Proc. ACM SIGKDD, 2002.

[16] T. Joachims and F. Radlinski, "Search Engines That Learn from Implicit Feedback," IEEE Computer Society, vol. 40, no. 8, pp. 34-40, 2007.

[17] Q. Tan, X. Chai, W. Ng, and D.L. Lee, "Applying Co-Training to Click through Data for Search Engine Adaptation," Proc. Ninth Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2004.

[18] F. Liu, C. Yu, and W. Meng, "Personalized Web Search for Improving Retrieval Effectiveness," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1, pp. 28-40, Jan. 2004.

[19] L. Deng, W. Ng, X. Chai, and D.L. Lee, "Spying Out Accurate User Preferences for Search Engine Adaptation," Advances in Web Mining and Web Usage Analysis, LNCS 3932, pp. 87-103, 2006.

[20] R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. EDBT Workshop, vol. 3268, pp. 588-596, 2004.

[21] S. Chuang and L. Chien, "Automatic Query Taxonomy Generation for Information Retrieval Applications," Online Information Rev., vol. 27, no. 4, pp. 243-255, 2003.

[22] Z. Dou, R. Song, and J.R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. 16th Int'l World Wide Web Conf. (WWW), 2007.

[23] B. Koester, "Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies," Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2016.

[24] Jiawei Liu, Jiawei Liu, Qingshan Li, Yishuai Lin and Yingjian Li, "A Query Suggestion Method Based on Random Walk and Topic Concepts" IEEE Computer Society, ICIS 2017, May-2017

[25] Asri Maspupah and Saiful Akbar, "T ext-based search engine application framework", International Conference on Data and Software Engineering, IEEE 2016

[26] Hoang Giang, Thi Thanh Sang Nguyen, "Transformed Document Modeling for Efficiently Searching ", Seventh International Confrence on Information Science and Technology, IEEE, 2017